فصلنامه آینده‌پژوهی مدیریت
سال بیست و چهارم/ شماره ۹۸ / بهار و تابستان ۱۳۹۲

بکارگیری تکنیک‌های خوشه‌بندی و الگوریتم ژنتیک در بهینه‌سازی درختان تصمیم گیری برای اعتبارسنجی مشتریان بانک ها

محمود البرزی
استادیار و عضو هیئت علمی دانشگاه آزاد اسلامی واحد علوم و تحقیقات
mahmood_alborzi@yahoo.com

محمد خان‌بابایی
عضو باشگاه پژوهشگران جوان، دانشگاه آزاد اسلامی، واحد علوم و تحقیقات (مسئول مکاتبات)
mohammadkhanbababaei@yahoo.com

محمدابراهیم محمدپور زرندی
دانشیار دانشگاه آزاد اسلامی واحد تهران مرکز ی
pourzarandi@yahoo.com

تاریخ دریافت: ۹۱/۵/۱۸ **تاریخ پذیرش:** ۹۱/۸/۲۱

چکیده

درختان تصمیم گیری به عنوان یکی از تکنیک های داده کاوی کاربرد زیادی در اعتبارسنجی مشتریان بانک و شناسایی آن ها برای اعطای تسهیلات اعتباری دارد. مسئله اصلی در پیچیدگی درختان تصمیم گیری، اندازه بیش از حد، عدم انعطاف پذیری و دقت کم در طبقه بندی است. هدف از این مقاله ارائه مدل ترکیبی در بهینه سازی درختان تصمیم گیری توسط تکنیک الگوریتم ژنتیک به منظور حل مسائل ذکر شده در فوق برای اعتبارسنجی مشتریان بانک است. به نظر می رسد بتوان با انتخاب ویژگی های مناسب و ساخت درختان تصمیم گیری توسط الگوریتم ژنتیک به کاهش پیچیدگی و افزایش انعطاف پذیری درختان تصمیم گیری پرداخت. در مدل ترکیبی پیشنهادی ابتدا داده های اعتباری توسط تکنیک خوشه بندی SimpleKmeans به دو خوشه تقسیم می شوند. سپس با استفاده از الگوریتم ژنتیک، پنج الگوریتم انتخاب ویژگی مبتنی بر سه رویکرد فیلتر، Wrapper و طرح جاسازی شده بر پایه درخت تصمیم گیری ژنتیکی، به انتخاب ویژگی های اعتبارسنجی مهم در مجموعه داده می پردازند. در ادامه پنج درخت تصمیم گیری مبتنی بر الگوریتم C4.5 در هر خوشه با مجموعه ویژگی های منتخب ساخته می شود. بهترین درختان تصمیم گیری در هر خوشه مبتنی بر معیارهای بهینگی مورد نظر در این مقاله انتخاب شده و با هم ترکیب می شوند تا درخت تصمیم گیری نهایی برای اعتبارسنجی مشتریان بانک ایجاد شود. ابزار یادگیری ماشین وکا و نرم افزار GATree برای رسیدن به نتایج بکار گرفته شده است. نتایج پژوهش نشان می دهد که استفاده از مدل ترکیبی پیشنهادی در ساخت درخت تصمیم گیری منجر به افزایش دقت طبقه بندی نسبت به بسیاری از الگوریتم های مقایسه شده در این مقاله می شود؛ ولی پیچیدگی الگوریتم مدل ترکیبی پیشنهادی از برخی الگوریتم های طبقه بندی مقایسه شده در این مقاله بیشتر است.

واژه‌های کلیدی: اعتبارسنجی، طبقه بندی، الگوریتم ژنتیک، درختان تصمیم گیری، انتخاب ویژگی، خوشه بندی.

۱- مقدمه

امروزه بانک ها برای شناخت مشتریان، ارضای نیازمندی‌ها و ارائه خدمات مالی مناسب نیازمند شناسایی دقیق ویژگی های اعتباری آن ها هستند. یکی از خدمات مالی در بانک، اعطای تسهیلات مالی از جمله وام به متقاضیان اعتباری است. بانک‌ها همچون کسب و کارهای دیگر در طول حیات خود با ریسک هایی مواجه می شوند. یکی از مهم ترین آن ها ریسک اعتباری است که باید با آن مقابله کنند. اعتبارسنجی[۱] به عنوان یک تکنیک موثر به شناخت مشتریان خوب و بد پرداخته و با این کار می تواند ریسک اعتباری آن ها را تعیین کند. تحقیقات زیادی بر روی مدل های اعتبارسنجی در بانک ها صورت گرفته است. در ابتدا مدل های اعتبارسنجی به صورت قضاوتی بودند. سپس روش های پارامتریک در اعتبارسنجی مطرح شدند. اخیرا از روش های ناپارامتریک در اعتبارسنجی مشتریان بانک‌ها استفاده می شود. فیشر[۲] در سال ۱۹۶۳ برای اولین بار ایده متمایز کردن گروه ها را مطرح کرد. سپس دیوید دراند[۳] در سال ۱۹۴۱ با تفکیک مشتریان به دو گروه خوب و بد به اعطای وام به آن ها پرداخت. کارت های اعتباری در سال ۱۹۶۰ وارد بانک‌ها شدند. در سال ۱۹۸۰ برای اولین بار از اعتبارسنجی در بانک‌ها استفاده شد. همچنین اعتبارسنجی در بازاریابی مستقیم در سال ۱۹۹۰ بکار رفت. موفقیت های اخیر در بکارگیری اعتبارسنجی منجر به استفاده از این تکنیک در ارزیابی اعتبار مشتریان شد(Thomas, 2000). در ادامه در بخش ۲ به بیان مسئله و ضرورت و اهداف پژوهش، بخش ۳ مروری بر تحقیقات صورت گرفته، بخش ۴ مواد و روش ها، بخش ۵ آموزش و تست مدل، بخش ۶ به مقایسه درخت تصمیم گیری حاصل از مدل ترکیبی پیشنهادی با سایر درختان تصمیم گیری و در نهایت در بخش ۷ به نتیجه گیری پرداخته می شود.

۲- بیان مسئله، ضرورت و اهداف تحقیق

تکنیک های داده کاوی همچون طبقه بندی می توانند با ارائه یک الگو یا مدل به کشف دانش پنهان در حجم زیادی از داده های تراکنش های اعتباری مشتریان بانک‌ها کمک کنند. درختان تصمیم گیری به عنوان یکی از تکنیک های طبقه بندی از روش های ناپارامتریک در اعتبارسنجی هستند. درختان تصمیم گیری می توانند با شناسایی ویژگی های مشتریان و تفکیک آن ها به گروه های خوب و بد، به اعتبارسنجی آن ها بپردازند. این ویژگی ها با توصیف مشخصات مشتریان بانک ها، طبقه آن ها را در اعتبارسنجی مشخص می کنند. از طرف دیگر برای شناخت الگوی مناسب در طبقه بندی مشتریان بانک در حجم زیادی از داده نیاز به پیش پردازش داده ها است. روش های مختلفی برای پیش پردازش داده ها وجود دارد که دو مورد از آن ها که در این مقاله بکار می روند، خوشه بندی و انتخاب ویژگی ها است. با پیش پردازش داده‌ها الگوی بهتری برای شناسایی و اعتبارسنجی مشتریان بانک ها ایجاد می شود.

مسئله اصلی در تحقیق موضوع این مقاله ساخت درختان تصمیم گیری است که بتوانند به طور بهینه به طبقه‌بندی مشتریان خوب و بد بانک‌ها برای اعتبارسنجی بپردازند. به نظر می رسد استفاده از الگوریتم های ژنتیک در انتخاب ویژگی ها و ساخت درختان تصمیم گیری بتواند منجر به طبقه بندی و اعتبارسنجی بهتری از مشتریان بانک ها شود. ممکن است الگوریتم های انتخاب ویژگی در بهینه محلی گیر کنند و همچنین تعامل بین ویژگی ها را در نظر نگیرند. در برخی از الگوریتم های انتخاب ویژگی فرض بر این است که روابط بین ویژگی ها خطی بوده و مستقل از هم می باشند. الگوریتم های انتخاب ویژگی تنها از برخی معیارها برای انتخاب ویژگی ها استفاده می کنند.

از طرف دیگر با توجه به اینکه الگوریتم درخت تصمیم گیری C4.5 یک الگوریتم بازگشتی[۴] و حریصانه[۵] است، منجر به ایجاد درخت تصمیم گیری پیچیده و بزرگ می شود. به همین دلیل قسمت های پایینی درخت تعداد کمی تراکنش را پوشش می دهند. تعداد کم تراکنش در قسمت های پایینی درخت اگرچه منجر به افزایش دقت درخت در طبقه بندی مشتریان می شود ولی از طرف دیگر انعطاف پذیری را در طبقه بندی مشتریان کاهش داده و در قسمت های پایینی درخت تنها با تغییر کوچک و تقریبا نامحسوس کلاس یا طبقه

یک مشتری جدید تغییر می کند. به همین دلیل به نظر می رسد اگر بتوان تا جایی که لطمه ای به دقت طبقه بندی در درخت تصمیم گیری نخورد، اندازه درخت، تعداد برگ ها و در نتیجه پیچیدگی درخت تصمیم گیری را کاهش داد، علاوه بر افزایش انعطاف پذیری در طبقه بندی، کاهش اندازه درخت و پیچیدگی آن می توان طبقه بندی بهتری را برای مشتریان بانک انجام داد.

در این مقاله فرض بر این است که بکارگیری الگوریتم ژنتیک در ساخت درختان تصمیم گیری مبتنی بر الگوریتم C4.5 منجر به بهینه سازی آن ها در دقت و پیچیدگی طبقه بندی مشتریان اعتبارسنجی بانک ها می شود. C4.5 یکی از الگوریتم های ساخت مدل طبقه بندی درخت تصمیم گیری است. این الگوریتم یک درخت تصمیم گیری به صورت نمودار گرافیکی ایجاد می کند که فهم آن برای کاربران آسان است. الگوریتم استنتاج C4.5 در سال ۱۹۹۳ توسط کوئینلن[۶] تهیه شد که از مفهوم شاخص کسب اطلاعات[۷] در ساخت درخت تصمیم گیری استفاده می کند(Larose, 2005). هدف از تحقیق موضوع این مقاله ارائه یک مدل مناسب برای اعتبارسنجی مشتریان بانک ها است. سوال تحقیق به صورت زیر است: چگونه می توان توسط الگوریتم ژنتیک به انتخاب ویژگی ها و ساخت درختان تصمیم گیری بهینه در اعتبارسنجی مشتریان بانک ها پرداخت؟

منظور از بهینگی درختان تصمیم گیری مدل پیشنهادی و سایر مدل های درخت تصمیم گیری مقایسه شده در این مقاله شامل موارد زیر است:

۱- تعداد ویژگی های پیشگویی کننده انتخابی[۸]؛ در یک درخت تصمیم گیری هر چه تعداد ویژگی های پیشگویی کننده کمتر باشد، پیچیدگی مدل کاهش یافته و انعطاف پذیری آن زیاد می شود. همچنین در بررسی وضعیت اعتباری متقاضیان اعتبار، زمان و هزینه کمتری صرف بررسی ویژگی های مشتریان می شود و از طرف دیگر سرعت پاسخگویی به مشتری به منظور رد یا تائید تقاضا بیشتر می شود. از طرف دیگر نیاز به ذخیره سازی داده های کمتری در سیستم پردازش مدل طبقه‌بندی و اعتبارسنجی مشتریان در بانک‌ها است.

۲- درصد مشاهدات درست طبقه بندی شده[۹] که حاصل تقسیم تعداد مشتریان درست طبقه بندی شده به تعداد کل مشتریان است.

۳- تعداد برگ ها در درخت تصمیم گیری. برگ در یک درخت تصمیم گیری گره انتهایی آن است که در این گره به تعیین خوب یا بد بودن مشتری اعتبارسنجی پرداخته می شود. افزایش تعداد برگ های درخت تصمیم گیری باعث افزایش پهنای درخت تصمیم گیری شده و پیچیدگی مدل اعتبارسنجی را افزایش می دهد.

۴- اندازه درخت تصمیم گیری که به تعداد شاخه ها و قوانین اگر آنگاه در درخت تصمیم گیری ارتباط دارد. اندازه درخت تصمیم گیری برابر با مجموع تعداد برگ ها و گره ها در آن است. با افزایش اندازه درخت تصمیم گیری پیچیدگی مدل اعتبارسنجی نیز افزایش می یابد.

با توجه به بیان مسئله و ضرورت تحقیق موضوع این مقاله، برخی از مزایایی که الگوریتم های ژنتیک می تواند در انتخاب ویژگی ها داشته باشند، شامل موارد زیر است (Carvalho & Freitas, 2004):

۱) الگوریتم های ژنتیک بر خلاف الگوریتم های حریصانه در یک لحظه با مجموعه ای از راه حل هل کار می کنند.

۲) الگوریتم های استنتاج حریصانه به بررسی تنها راه حل های جزئی در هر مرحله می پردازند.

۳) الگوریتم های ژنتیک با توجه به استفاده از قانون احتمال در بهینه محلی کمتر گیر می کنند. الگوریتم های ژنتیک به ایجاد یک رابطه مناسب بین اندازه و پیچیدگی درختان تصمیم گیری می پردازند (Aitkenhead, 2008). همچنین مزیت های بکارگیری الگوریتم ژنتیک در ساخت درختان تصمیم گیری می تواند شامل موارد زیر باشد:

۱) روش های تکاملی برای ایجاد درخت باعث می شوند که تغییرات به طور اتوماتیک در درخت ایجاد شوند. می توان توسط ترکیب روش های تکاملی و درخت تصمیم بین اندازه درخت و پیچیدگی آن رابطه مناسبی ایجاد نمود(Aitkenhead, 2008)

۲) انعطاف پذیری در نمایش تابع طبقه بندی.

۳) می توان توابع برازندگی متنوعی در طبقـه بنـدی بوجود آورد(Zhang & Bhattacharyya, 2004)

۳- مروری بر تحقیقات صورت گرفته

در مورد کاربرد مدل های پارامتریک و ناپارامتریک اعتبارسنجی در طبقه بندی مشتریان بانک ها و موسسات مالی تحقیقات متنوعی در داخل و خارج از کشور صورت گرفته است. جدول ۱ به چند نمونه از این تحقیقات در خارج از کشور اشاره می کند.

در ادامه به تحقیقاتی راجع به بکارگیری الگوریتم های ژنتیک در انتخاب ویژگی ها در جدول ۲ و ساخت درختان تصمیم گیری در جدول ۳ اشاره می شود. این تحقیقات در حوزه های دیگر علوم و کسب و کارها می باشند و می توان از نتایج آن ها در اعتبارسنجی مشتریان بانک ها استفاده نمود.

بکارگیری الگوریتم ژنتیک در انتخاب ویژگی ها در سال ۱۹۸۹ و اولین بار به وسیله Siedlecki و Sklansky در مقاله ای تحت عنوان A note on genetic algorithms for large-scale feature selection مطرح شد (Tsang, Kwong, & Wang, 2007).

جدول ۱- برخی تحقیقات روش های اعتبارسنجی در خارج از کشور

ردیف	مدل اعتبارسنجی	پژوهشگر
۱	رگرسیون لجستیک، شبکه عصبی، درخت تصمیم گیری	(Susac, Sarlija, & Bensic, n.d.)
۲	ترکیب تحلیل تمایزی و الگوریتم پس انتشار در شبکه عصبی	(Lee, Chiu, Lu, & Chen, 2002)
۳	الگوهای طبقه بندی غلط	(Kim & Sohn, 2004)
۴	ترکیب مدل های شبکه های عصبی مصنوعی و روش MARS	(Lee & Chen, 2005)
۵	رتبه بندی تحلیل لینک با استفاده از ماشین بردار پشتیبان	(Xu, Zhou, & Wang, 2008)
۶	شبکه های عصبی و تکنیک های عمومی	(Abdou & Pointon, 2008)
۷	طبقه کننده های ترکیبی به جای یک طبقه کننده	(Nanni & Lumini, 2009)

جدول ۲- برخی تحقیقات در کاربرد الگوریتم ژنتیک در انتخاب ویژگی ها

ردیف	رویکرد بکارگیری الگوریتم ژنتیک در انتخاب ویژگی ها	پژوهشگر
۱	الگوریتم REDUCE	(Lavrac, Gamberger, & Turney, 1995)
۲	استفاده از عملگر خودتقاطعی برای انتخاب ویژگی ها	(PAL, NAND, & Kundu, 1998)
۳	ترکیب طبقه بندی چندگانه مبتنی بر الگوریتم ژنتیک	(Lee,2002 cited in &Kim ,Kim خان بابایی & نادعلی،۱۳۸۷)
۴	استفاده از الگوریتم ژنتیک برای انتخاب متغیرهای ورودی	(D'heygere, Goethals, & Pauw, 2003)
۵	الگوریتم ژنتیک در انتخاب متغیرها به کمک خوشه بندی مشتریان	(Liu & Ong, 2008)
۶	استفاده از الگوریتم ژنتیک در ترکیب روش های انتخاب ویژگی	(Tan, Fu, Zhang, & Bourgeois, 2008)

جدول ۳- برخی تحقیقات در کاربرد الگوریتم ژنتیک در ساخت درختان تصمیم گیری

ردیف	رویکرد بکارگیری الگوریتم ژنتیک در ساخت درختان تصمیم گیری	پژوهشگر
۱	ساخت درختان تصمیم گیری دودویی توسط تکنیک های تکاملی الگوریتم ژنتیک	(Papagelis & Kalles, n.d.)
۲	ترکیب الگوریتم های ژنتیک و درختان تصمیم ID3 در طبقه بندی الگو	(Bala, Huang, Vafaie, DeJong, & Wechsler, 1995)
۳	استفاده از الگوریتم ژنتیک برای کشف قوانین در شقوق کوچک درختان تصمیم گیری	(Freitas & Carvalho, ۲۰۰۲،۲۰۰۴)
۴	یکپارچه کردن الگوریتم قوانین وابستگی و الگوریتم ژنتیک درکشف درخت طبقه بندی	(Hsu, & Lai, Chiu ,Hsu خان بابایی & نادعلی،۱۳۸۷ 2003 cited in)
۵	کشف قوانین تصمیم گیری در ورشکستگی به کمک تصمیمات کیفی خبرگان	(Kim & Han, 2003)
۶	استفاده از الگوریتم های ژنتیک برای یادگیری سلسله مراتب طبقه کننده ها	(Martinez-Otzeta, Sierra, Lazkano, & Astigarraga, 2006)
۷	بهنیه سازی پیشگویی مدل ها بر مبنای درختان تصمیم و شبکه های عصبی	(D'heygere, Goethals, & Pauw, 2006)

ردیف	رویکرد بکارگیری الگوریتم ژنتیک در ساخت درختان تصمیم گیری	پژوهشگر
۸	بررسی سودمندی تکنیک درخت تصمیم گیری بر مبنای الگوریتم ژنتیک	(Huang, Gong, Shi, Liu, & Zhang, 2007)
۹	تعریف برازندگی درخت توسط الگوریتم ژنتیک	(نادعلی ,۱۳۸۷ & خان بابایی Janssens, & Sorensen (2003 cited in
۱۰	تحلیل درخت طبقه بندی توسط الگوریتم TARGET	(Gray & Fan, 2008)
۱۱	الگوریتم ژنتیک چند هدفه Elitist به کشف قوانین طبقه بندی مجموعه داده های بزرگ	(Dehuri, Patnaik, Ghosh, & Mall, 2008)

۴- مواد و روش ها

۴-۱- مجموعه داده

داده هایی که در این مقاله برای ساخت و آزمون درختان تصمیم گیری C4.5 مورد استفاده قرار می گیرند، مجموعه داده های اعتباری آلمان[۱۰] است(http://archive.ics.uci.edu/ml/datasets.html) که در مقالات مشابه برای بررسی اثربخشی و امکان سنجی مدل مورد استفاده قرار گرفته است. این مجموعه داده که در سال ۱۹۹۲ تهیه شد، فاقد مقادیر مفقود و اختلال است. روی این مجموعه عملیات آماده سازی و تمیز کردن و پیش پردازش داده ها صورت می گیرد. این مجموعه داده دارای ۱۰۰۰ تراکنش و ۲۱ ویژگی است. از این تعداد ویژگی ۷ ویژگی عددی و ۱۳ تای آن اسمی هستند. یک ویژگی هدف در این ویژگی ها به بررسی خوب یا بد بودن مشتری می پردازد. ویژگی های مجموعه داده های اعتباری آلمان به همراه مقادیر و نوع آن ها در پیوست ۱ موجود است.

با انجام فرایند آماده سازی و تمیز کردن داده ها در چند مرحله و با اعمال روش های مختلف توسط نرم افزار یادگیری ماشین وکا[۱۱] نسخه ۳,۵,۸، تعداد تراکنش ها از ۱۰۰۰ به ۶۹۰ کاهش یافت. همچنین همه ویژگی ها از نوع عددی به نوع اسمی تبدیل شدند. توسط نمودارهای قابلیت تجسم سازی[۱۲] موجود در این نرم افزار، وضعیت کلی داده ها در هر ویژگی، مشاهده و ارتباط بین مقادیر هر یک از ویژگی های پیشگویی کننده با مقادیر ویژگی هدف بررسی شد. روش های زیر در آماده سازی داده ها برای آموزش و تست مدل ترکیبی پیشنهادی برای طبقه بندی مشتریان اعتباری از اولویت بندی خاصی پیروی نمی کند. برخی روش ها، چندین بار بر روی مجموعه داده در مراحل مختلف آماده سازی اعمال شده اند. روش های آماده سازی داده ها در این مقاله بدین صورت است:

۱. حذف مقادیر پراکنده. ۲. نرمال سازی که فقط بر روی ویژگی "سن" اعمال شد و همچنین مقادیر ویژگی هدف در محاسبات این روش در نظر گرفته شد. ۳. گسسته سازی[۱۳] مقادیر ویژگی های عددی که در این روش مقادیر ویژگی هدف در محاسبات لحاظ شد. ۴. ادغام مقادیر[۱۴] داده در ویژگی های اسمی. ۵. تبدیل[۱۵] ویژگی های عددی به اسمی.

۴-۲- ریسک اعتباری و اعتبارسنجی

بر اساس دیدگاه کمیته بال[۱۶] مهم ترین ریسک هایی که بانک ها با آن مواجه هستند شامل موارد زیر است: ریسک اعتباری، ریسک کشوری و ریسک انتقال وجوه، ریسک بازار، ریسک نرخ بهره، ریسک نقدینگی، ریسک عملیاتی، ریسک حقوقی و ریسک شهرت (اداره مطالعات و کنترل ریسک بانک تجارت, ۱۳۸۶). بانک ها برای مقابله با ریسک های پیش رو باید بتوانند به مدیریت ریسک بپردازند. یکی از ریسک های بانک ها ریسک اعتباری است. ریسک اعتباری به معنی احتمال عدم بازپرداخت اصل و سود تسهیلات اعطایی توسط گیرنده اعتبار به علت عدم تمایل و یا ناتوانی مالی است(اداره مطالعات و کنترل ریسک بانک تجارت, ۱۳۸۶). برای هر موسسه ارائه دهنده خدمات مالی مثل بانک های تجاری، توانایی تفکیک مشتریان خوب و مشتریان بد، امری حیاتی و مهم می باشد. در نتیجه نیاز به مدل های معتبری است که توسط آن ها بتوان به پیش بینی قصور در بازپرداخت وام توسط مشتریان پرداخت تا ذینفعان بتوانند در زمان مناسب اقدامات پیشگیرانه و اصلاحی صحیحی انجام دهند(Yu, Wang, & Lai, 2007) .

یکی از تکنیک های مهم در ارزیابی ریسک اعتباری، اعتبارسنجی است. توماس[۱۷] در سال ۲۰۰۲ اعتبارسنجی را تکنیکی تعریف کرد که به بانک ها و شرکت های

اعتباری در زمینه اعطای اعتبار به مشتریان بر مبنای معیارهای از قبل تعیین شده، کمک می کند(Yu, Wang, & Lai, 2007). اعتبارسنجی دارای مزیت هایی مثل موارد زیر می باشد: ۱. کاهش هزینه تحلیل اعتبار. ۲. تصمیم گیری سریع در اعتبارسنجی مشتریان. ۳. تضمین اعتبارات و حذف ریسک های احتمالی (Nanni & Lumini, 2009) (Ong, Huang, & Tzeng, 2005). بریل[۱۸] در سال ۱۹۹۸ علاوه بر دو مورد بالا به موارد زیر نیز اشاره کرد(1387, نادعلی & خان بابایی, Huang, Chen, & Wang, 2007 cited in): ۱. نظارت نزدیک به حساب های موجود. ۲. تعیین اولویت در مجموعه اعتبارات.

روش های اعتبارسنجی در ابتدا قضاوتی بود. کارشناس اعتبارسنجی با بررسی فرم تقاضانامه مبتنی بر تحلیل پنج ضابطه معروف به تحلیل (پنج C) به تصمیم گیری در مورد اعطای وام یا رد تقاضانامه می پرداخت. 5C برگرفته از حروف اول ۵ کلمه است. این ۵ کلمه عبارتند از: ویژگی[۱۹] شخص گیرنده وام، سرمایه[۲۰] وی، ضمانت[۲۱]، توانایی[۲۲] بازپرداخت و شرایط[۲۳] (Thomas, 2000). امروزه اکثر تحقیقات و کاربردها حول محور اعتبارسنجی توسط دو روش صورت می گیرند: ۱. روش های سنتی مثل رگرسیون لجستیک و مدل های لوجیت و پروبیت. ۲. روش های داده کاوی (Sabzevari, Soleymani, & Noorbakhsh, n.d.). در تقسیم بندی دیگری مدل های اعتبارسنجی به مدل های پارامتریک و ناپارامتریک تقسیم می شوند (Sabzevari, Soleymani, & Noorbakhsh, n.d.). مدل های پارامتریک مثل تحلیل تمایزی، رگرسیون خطی، پروبیت و لوجیت و مدل های ناپارامتریک مثل درختان طبقه بندی، شبکه های عصبی، سیستم های خبره و ...

بسیاری از تحقیقات اشاره شده در بخش ۳ این مقاله از روش های پارامتریک و ناپارامتریک برای اعتبارسنجی مشتریان استفاده کردند. نکته قابل توجه در این تحقیقات این است که هر یک از مدل های موجود در اعتبارسنجی به شیوه ای خاص به متمایز کردن مشتریان خوب و بد پرداختند. در این تحقیقات هر یک از مدل ها بر دیگری برتری داشتند و در سال های اخیر نیز محققین به این قضیه توجه خاصی کرده اند. آن ها سعی دارند به شیوه ای بهتر به طبقه بندی مشتریان اعتبارسنجی بپردازند. دقت، پیچیدگی و انعطاف پذیری در طبقه بندی مورد توجه آن ها در مقایسه مدل های اعتبارسنجی است. در حال حاضر روش های جدیدی از هوش مصنوعی نیز در اعتبارسنجی بکار رفته اند که از آن ها می توان به موارد زیر اشاره کرد: شبکه های عصبی مصنوعی، محاسبات تکاملی، الگوریتم های ژنتیک و ماشین بردار پشتیبان. همچنین اخیرا از مدل های ترکیبی در اعتبارسنجی استفاده می شود؛ مثل مدل عصبی فازی، مدل فازی ماشین بردار پشتیبان و مدل ترکیبی شبکه های عصبی(Yu, Wang, & Lai, 2007). در این مقاله از ترکیب روش های خوشه بندی، الگوریتم انتخاب ویژگی ها مبتنی بر الگوریتم های ژنتیک، درختان تصمیم گیری مبتنی بر الگوریتم ژنتیک و درختان تصمیم گیری C4.5 برای ارائه مدل ترکیبی پیشنهادی اعتبارسنجی مشتریان بانک ها استفاده می شود. این روش ها از نوع روش های داده کاوی، ناپارامتریک، هوش مصنوعی و مدل های ترکیبی هستند.

۴-۳- طبقه بندی و درختان تصمیم گیری

یکی از وظایف داده کاوی طبقه بندی است. طبقه بندی دارای تکنیک های متنوعی است که در تحقیقات مختلف از آن ها استفاده می شود. از تکنیک های رایج در طبقه بندی می توان به موارد زیر اشاره کرد: K نزدیک ترین همسایه، درختان تصمیم گیری، شبکه های عصبی، ماشین بردار پشتیبان، طبقه بندی بیزین، رگرسیون، تئوری های مجموعه دانه درشت، منطق یابی مبتنی بر حالت، سیستم های خبره، منطق فازی، الگوریتم های ژنتیک. درختان تصمیم گیری به علت سادگی و قابلیت فهم بالا از محبوبیت بالایی در کاربرد برخوردار هستند. این تکنیک در زمره درختان طبقه بندی قرار می گیرد. درختان طبقه بندی به پیشگویی مقادیر ویژگی ها یا متغیرهای وابسته و گسسته می پردازد. درختان تصمیم گیری تنها مقادیر ویژگی های گسسته را پیشگویی می کنند. این پیشگویی توسط متغیر کلاس که ویژگی هدف یا ویژگی وابسته نیز نامیده می شود، صورت می گیرد.

مقادیر ویژگی هدف، وابسته به مقـادیر متغیرهـای(ویژگی های) مستقل(توصیف کننده) و وجود آن هـا در سـاختار درخت تصمیم گیـری اسـت (D'heygere, Goethals, & Pauw, 2003). درختـان تصـمیم گیـری دارای الگـوریتم های مختلفی هستند که برخی از آن ها شامل موارد زیـر است: ID3، C4، C4.5، C5، CART، CHAID، QUEST. در این مقاله از الگوریتم درخت تصمیم گیری C4.5 برای ساخت درختان تصمیم گیری بـه منظـور طبقـه بنـدی و اعتبارسنجی مشتریان بانک ها استفاده می شود.

الگـوریتم اسـتنتاج C4.5 در سـال ۱۹۹۳ توسـط کوئینلن تهیه شـد. ایـن الگـوریتم متغیرهـای پیوسـته و گسسته را در محاسبات خود لحاظ کرده و مقادیر مفقـود را در الگـوریتم خـود در نظـر مـی گیـرد (Aitkenhead, 2008). این الگوریتم لزوما دودویی نیست. بـرای انتخـاب یک جداکننده بهینه در طول مسیر درخت تصمیم گیری از شاخص کسب اطلاعات یا کـاهش آنتروپـی[۲۴] اسـتفاده مـی کنـد (Larose, 2005). بـرای فهـم شـاخص کسـب اطلاعات، آنتروپی و شاخص کسب می توان بـه (Tsang, Kwong, & Wang, 2007) رجـوع کـرد و الگـوریتم اسـتنتاج C4.5 در (Larose, 2005) آمـده اسـت. ایـن الگوریتم نسخه جدید الگوریتمID3 است. در این مقاله بـه چند دلیل از الگـوریتم درخـت تصـمیم گیـری C4.5 بـه جای الگوریتم ID3 استفاده می شود:

۱. الگوریتم ID3 تنها ویژگی های اسمی را در ساخت درخت تصمیم گیری در نظر مـی گیـرد، ولـی الگـوریتم C4.5 هر دو نوع ویژگی اسمی و عددی را لحاظ می کند.

۲. در الگوریتم ID3 ابتدا بایـد داده هـای مفقـود را از بین برد، ولی C4.5 در الگوریتم خود با داده هـای مفقـود مقابله می کند.

۳. الگـوریتم C4.5 بـر خـلاف ID3 بـه هـرس کـردن درخت مـی پـردازد (Hall, 1999). هـرس درخـت باعـث کاهش اندازه درخت و پیچیدگی آن مـی شـود. الگـوریتم درخت تصمیم گیری C4.5 برای هر مقدار ویژگی اسـمی به طور پیش فرض به تفکیک شاخه می پردازد که موجب پرپشت شدن[۲۵] درخت تصمیم گیری می شود (Larose, 2005).

۴-۴- خوشه بندی

خوشه بندی به عنوان یکی از فعالیت های داده کاوی می باشد و به گروه بندی کردن تراکنش ها، مشاهدات یـا حالت ها در کلاس های مشابه مـی پـردازد. یـک خوشـه مجموعه ای از رکوردها است که به هم شبیه می باشند و از رکوردهای بیرون خوشه تفاوت دارند. در خوشه بنـدی متغیر هدف وجـود نـدارد و بـه طبقـه بنـدی، تخمـین و پیشگویی مقدار متغیر هدف نمی پردازد(Larose, 2005) . در این مقاله از الگـوریتم خوشـه بنـدی SimpleKmeans استفاده می شود. معیار نزدیکـی در پیـدا کـردن نزدیـک ترین مرکز خوشه برای هر رکورد، معمولا فاصله اقلیدسی است. معیار توقف می تواند به طور مثال مجموع مربعـات خطا باشد. الگوریتم SimpleKmeans در (Olson & Shi, 2007,p.75) آمده اسـت. ۱. انتخـاب تعـداد مـورد تمایـل خوشه ها به اندازه K. ۲. انتخاب تعداد K مشـاهده اولیـه به عنوان seed. ۳. محاسبه متوسط مقادیر خوشه برای هر ویژگی یا متغیر. ۴. تخصیص مشاهدات آموزشـی[۲۶] دیگـر به نزدیک ترین خوشه توسط محاسبه مقیاس فاصله مورد نظر. ۵. محاسبه مجدد متوسـط هـای خوشـه بـر اسـاس تخصیص ها در مرحله ۴. ۶. تکرار بین مراحل ۴ و ۵. مـی توان از تکنیک خوشه بندی به عنوان پیش پردازش داده ها استفاده کرد(Olson & Shi, 2007) کـه در ایـن مقالـه این تکنیک بر روی مجموعـه داده هـای اعتبـاری آلمـان اعمال می شود.

۴-۵- انتخاب ویژگی ها

در مدل ناپارمتریک طبقه بندی هزینه و زمان زیـادی باید صرف کسب داده های مـدل شـود؛ زیـرا مـدل هـای ناپارامتریک مبتنی بر داده هسـتند. پـس بایـد بـه جمـع آوری ویژگـی هـا و داده هـایی پرداخـت کـه از اهمیـت بیشتری برای ساخت مدل طبقه بندی برخوردارند. حذف اطلاعات غیر مـرتبط و اسـتخراج متغیرهـای کلیـدی در شناخت الگو، پیش پردازش نامیده می شود. (Kennedy, Lee, Roy, Reed, & Lippmann, 1998). بـرای سـاخت یک مدل طبقه بندی مناسب نیاز به داده های آموزشی با کیفیت است. این کیفیت با تعداد داده ها و ویژگی هـا در مجموعه آموزش ارتباط دارد. انتخاب ویژگی ها به عنـوان

یکی از روش های پیش پردازش داده می تواند باعث افزایش کیفیت مجموعه داده آموزش برای ساخت مدل طبقه بندی گردد. درختان تصمیم گیری به عنوان یکی از تکنیک های طبقه بندی نیز از این قاعده مستثنی نیستند (SALAPPA, DOUMPOS, & ZOPOUNIDIS, 2007). در این مقاله از الگوریتم های انتخاب ویژگی به عنوان یکی از روش ها برای پیش پردازش داده ها استفاده می شود.

تعاریف مختلفی از انتخاب ویژگی ها مطرح شده است. انتخاب ویژگی به شناسایی و انتخاب ویژگی های متمایز برای ساخت مدل ها و تفسیر بهتر داده ها می پردازد(Tan, Fu, Zhang, & Bourgeois, 2008) . انتخاب ویژگی ها دارای مزایای متعددی است: ۱. باعث فهم آسان داده ها می شود. ۲. زمان یادگیری را در مدل کاهش می دهد. ۳. با انتخاب ویژگی ها نیاز کمتری به اندازه گیری و ذخیره سازی مقادیر ویژگی ها است (Guyon & Elisseeff, 2003).

با توجه به این موضوع می توان گفت که انتخاب ویژگی ها باعث می شود یک مدل اعتبارسنجی بهتری برای طبقه بندی مشتریان بانک ها تولید شود و از طرف دیگر در حجم زیاد داده های اعتبارسنجی، هزینه و زمان جمع آوری و بررسی ویژگی های مشتریان جدید کاهش یابد. در نتیجه کارشناسان اعتبارسنجی می توانند سریعتر به تصمیم گیری در مورد قبول یا رد تقاضای اعتبارسنجی مشتریان بانک ها بپردازند. همچنین با آسان شدن فهم داده ها، تفسیر رد یا قبول اعتبار برای کارشناسان اعتبارسنجی و متقاضیان اعتبار راحت تر می شود.

الگوریتم انتخاب ویژگی ها از سه قسمت تشکیل می شود(WANG & LI, 2008):

۱. معیار ارزیابی ویژگی.
۲. روش جستجو.
۳. قانون توقف.

به طور معمول معیارهای ارزیابی شامل موارد زیر است: ۱. اطلاعات. ۲. وابستگی. ۳. فاصله. ۴. سازگاری. ۵. دقت طبقه بندی. الگوریتم های انتخاب ویژگی که از ۴ معیار اول اشاره شده در بالا استفاده می کنند مبتنی بر رویکرد فیلتر هستند. در این رویکرد، الگوریتم انتخاب ویژگی مستقل از الگوریتم طبقه بندی است. الگوریتم انتخاب ویژگی که از معیار دقت طبقه بندی برای انتخاب ویژگی ها استفاده می کند، از رویکرد Wrapper بهره می برد. در این رویکرد الگوریتم انتخاب ویژگی از الگوریتم یادگیری مثل الگوریتم طبقه بندی برای انتخاب ویژگی ها استفاده می کند. روش های جستجو در انتخاب ویژگی ها شامل ۳ روش است که عبارتند از: ۱. کامل. ۲. هیوریستک. ۳. تصادفی. دو روش کامل و هیوریستیک مربوط به فضاهای کوچک است و در مواردی مناسب می باشد که نیاز به کارایی بالا در فرایند جستجو می باشد. روش تصادفی مثل الگوریتم ژنتیک برای فضاهای بزرگ و پیچیده مناسب تر است. قوانین مختلفی برای توقف الگوریتم انتخاب ویژگی ها موجود است: ماکزیمم تعداد تکرار الگوریتم، کسب نتیجه بهتر توسط اضافه یا کم کردن یک ویژگی از مجموعه ویژگی ها، رسیدن به یک زیرمجموعه بهینه از ویژگی ها و ...(WANG & LI, 2008) . یکی دیگر از روش های انتخاب ویژگی، طرح های جاسازی شده است. در این روش الگوریتم انتخاب ویژگی به عنوان بخشی از الگوریتم طبقه بندی لحاظ می شود (SALAPPA, DOUMPOS, & ZOPOUNIDIS, 2007).

در این مقاله از رویکردهای فیلتر، Wrapper و طرح جاسازی شده برای انتخاب ویژگی ها استفاده می شود. روش جستجو در انتخاب ویژگی ها به صورت تصادفی و مبتنی بر الگوریتم ژنتیک است و قانون توقف برای الگوریتم انتخاب ویژگی ها، ماکزیمم تعداد تکرار در الگوریتم انتخاب ویژگی می باشد. هر یک از الگوریتم های انتخاب ویژگی دارای یک تابع ارزیاب[۲۷] برای ارزیابی ویژگی ها هستند. سه الگوریتم اول از رویکرد فیلتر، الگوریتم چهارم از رویکرد Wrapper و الگوریتم آخر از رویکرد طرح جاسازی شده برای انتخاب ویژگی ها استفاده می کنند. در ادامه به توضیحی مختصر راجع به روش های انتخاب ویژگی در این مقاله پرداخته می شود.

۱) الگوریتم انتخاب ویژگی مبتنی بر الگوریتم ژنتیک

توسط تابع ارزیاب همبستگی بین ویژگی ها با هم و با ویژگی هدف[۲۸]: تابع ارزیاب در این الگوریتم به بررسی همبستگی تک تک ویژگی ها با هم و با

ویژگی هدف می پردازد. تابع ارزیاب در این الگوریتم، الگوریتمی است که با جستجوی هیوریستیک، به بررسی همبستگی بین ویژگی ها می پردازد. این تابع ارزیاب فرض می کند که بهترین ویژگی ها نسبت به هم همبستگی کمتر و نسبت به ویژگی هدف دارای همبستگی بیشتری می باشند. البته همبستگی در این تابع ارزیاب به وابستگی ویژگی ها با هم دلالت دارد و منظور همبستگی خطی کلاسیک نیست(Hall, 1999) .

۲) الگوریتم انتخاب ویژگی مبتنی بر الگوریتم ژنتیک توسط تابع ارزیابی سازگاری زیر مجموعه ویژگی ها با مقادیر ویژگی هدف: به جستجوی کامل و جامع در فضای زیر مجموعه ویژگی ها می پردازد، تا اینکه کمینه ترین ترکیب از ویژگی ها را پیدا کند. سپس این مجموعه ویژگی ها به تقسیم مجموعه آموزشی در کلاس ها می پردازند. این الگوریتم توسط Liu و Setiono ابداع شد و ویژگی بارز آن این است که با اختلال در داده ها به خوبی برخورد می کند. این الگوریتم ابتدا به طور تصادفی یک زیر مجموعه با نام S از کل ویژگی ها انتخاب می کند. سپس در مرحله بعد یک زیر مجموعه دیگر از ویژگی ها تولید می شود. سطح سازگاری مقادیر ویژگی هدف توسط قرار دادن نمونه های مجموعه داده در این مجموعه ویژگی سنجیده می شود. اگر این سطح سازگاری کمتر باشد، این مجموعه جایگزین مجموعه قبلی می شود. نرخ ناسازگاری در مجموعه ویژگی ها در هر مرحله محاسبه می شود. این روند مرتبا ادامه می یابد تا مناسب ترین مجموعه ویژگی ها انتخاب شوند(Hall, 1999). برای فهم بیشتر می توان به (Liu&Setiono,n.d.) مراجعه کرد.

۳) الگوریتم انتخاب ویژگی مبتنی بر الگوریتم ژنتیک توسط تابع ارزیاب طبقه کننده C4.5: مجموعه ویژگی ها را توسط تابع طبقه کننده C4.5 در مجموعه داده آموزشی ارزیابی می کند. میزان شایستگی زیر مجموعه ویژگی منتخب توسط شاخص دقت طبقه بندی الگوریتم C4.5 تعیین می شود.

٤) الگوریتم انتخاب ویژگی مبتنی بر الگوریتم ژنتیک توسط تابع ارزیاب wrapper با طبقه کننده C4.5: توسط الگوریتم های یادگیری مثل الگوریتم C4.5، ویژگی ها را ارزیابی می کند. به عبارت دیگر الگوریتم انتخاب ویژگی، برای انتخاب ویژگی های مناسب از یک الگوریتم یادگیری بهره می برد. برای بکارگیری الگوریتم یادگیری در انتخاب ویژگی از داده های تست و آموزش در تکنیک اعتبارسنجی متقاطع استفاده می شود.

۵) الگوریتم انتخاب ویژگی مبتنی بر درخت تصمیم گیری ژنتیکی: این الگوریتم مبتنی بر طرح جاسازی شده در انتخاب ویژگی ها است. زیرا با اجرای این الگوریتم درخت تصمیم گیری ایجاد می شود که از مجموعه ویژگی های این درخت تصمیم گیری به عنوان ویژگی های منتخب در ساخت درخت تصمیم گیری C4.5 استفاده می شود. الگوریتم درخت تصمیم گیری ژنتیکی مورد استفاده در انتخاب ویژگی ها برگرفته از (Papagelis & Kalles, n.d.) است. در این الگوریتم برای نمایش راه حل ها از روش درختی استفاده می شود. همچنین در ساخت درختان تصمیم گیری به تغییرات اساسی در الگوریتم ژنتیک می پردازد. هر ویژگی دارای یک مقدار تصادفی می باشد. اگر ویژگی اسمی باشد، یکی از مقادیر آن در هر تکرار به طور تصادفی انتخاب می شود و اگر ویژگی مورد نظر عددی باشد، در بازه تعریف شده آن مقدار آن تغییر می کند. عملگر جهش ویژگی را به طور تصادفی انتخاب و مقدار آن را به طور تصادفی تغییر می دهد و عملگر تقاطع با انتخاب ویژگی ها به صورت تصادفی زیر درخت های آن ها را جابجا می کند. تابع برازندگی در اینجا به بهینه کردن اندازه درخت و دقت طبقه بندی می پردازد.

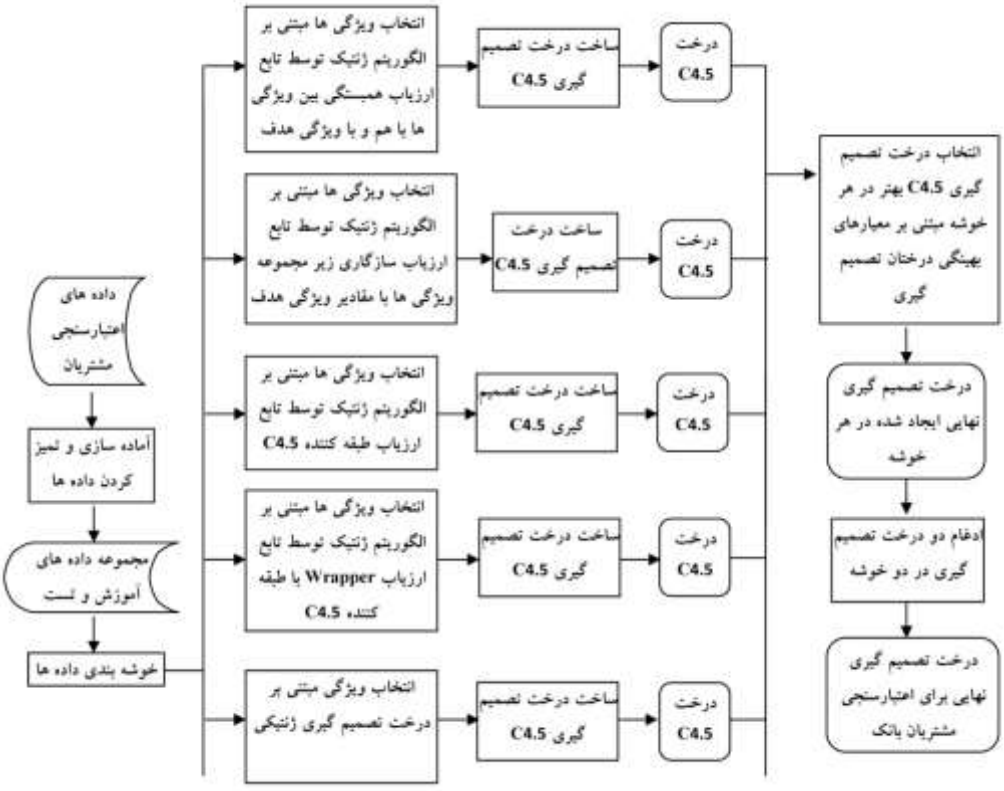
۴-۶- الگوریتم ژنتیک

واژه الگوریتم ژنتیک به توصیف یک مجموعه ای از رویه های جستجوی تصادفی می پردازد که از اصول ژنتیک طبیعی و اصل بقای برترین ها نشات گرفته شده

است. جان هلند[۲۹] اولین شخصی بود که در سال ۱۹۷۰ به طراحی الگوریتم ژنتیک پرداخت (1387, نادعلی & خان بابایی Grune & Jooste, 2004 cited in). از آن موقع تا حال تحقیقات فراوانی پیرامون این تکنیک و کاربردهای آن انجام شده است. مراحل اجرای الگوریتم ژنتیک می تواند شامل موارد زیر باشد(1387, نادعلی & خان بابایی Grune & Jooste, 2004 cited in): ۱. شناخت ژن ها: در الگوریتم ژنتیک برای نمایش یک ژن از یک بیت استفاده می شود. یک ژن نشان دهنده رفتار یک بخشی از راه حل مسئله است.۲. سازمان دهی ژن ها در کروموزوم ها : به مجموعه ژن ها کروموزوم می گویند. هر کروموزوم شامل مقادیری است که جمعا رفتار یک کروموزوم یا یک راه حل مسئله تبیین می کنند. ۳. ایجاد یک جمعیتی از راه حل های ممکن: به مجموعه ای از راه حل ها یا کروموزوم ها یک جمعیت می گویند.۴. ارزیابی تک تک کروموزوم ها: در این مرحله الگوریتم ژنتیک به ارزیابی اثربخشی جمعیت اولیه برای مقایسه با جمعیت های دیگر می پردازد.۵. ایجاد جمعیت جدید توسط عملیات تقاطع و جهش: برخی راه حل ها ممکن است در یک جمعیت از راه حل های دیگر برتر باشند. به همین دلیل با انتخاب این مجموعه راه حل ها و توسط عملیات تقاطع و جهش جمعیت دیگری شکل می گیرد. مجددا مراحل ۴ و ۵ انجام می شود تا شرط توقف الگوریتم پذیرفته شود.

۴-۷- مدل ترکیبی پیشنهادی

در مدل ترکیبی پیشنهادی تحقیق از الگوریتم های انتخاب ویژگی مبتنی بر الگوریتم ژنتیک، درخت تصمیم گیری ژنتیکی، درختان تصمیم گیری C4.5 و همچنین الگوریتم SimpleKmeans برای خوشه بندی داده ها استفاده می شود. شکل ۱ فرایند ساخت و آزمون مدل ترکیبی پیشنهادی موضوع این مقاله را برای اعتبارسنجی مشتریان بانک نشان می دهد.



شکل ۱ - فرایند ساخت و آزمون مدل ترکیبی پیشنهادی موضوع این مقاله در اعتبارسنجی مشتریان بانک

همان طور که مشاهده می شود، این مجموعه داده، بعد از عملیات آماده سازی و تمیز کردن به دو قسمت داده های آموزش و تست تقسیم می شود و به وسیله تکنیک های خوشه بندی و الگوریتم های انتخاب ویژگی، پیش پردازش داده ها صورت می گیرد. توسط تکنیک خوشه بندی SimpleKmeans این داده ها به دو دسته خوشه بندی می شوند. در ادامه در هر خوشه توسط پنج الگوریتم انتخاب ویژگی، ویژگی های مهم انتخاب شده و از این ویژگی ها تعداد پنج درخت تصمیم گیری J48)C4.5 در نرم افزار وکا) ساخته می شود. سپس به انتخاب بهترین درختان تصمیم گیری در هر خوشه مبتنی بر معیارهای بهینگی درختان تصمیم گیری مطرح شده در بخش ۲ این مقاله پرداخته می شود. بعد از اینکه درختان تصمیم گیری C4.5 برتر در هر دو خوشه تعیین شد، این دو درخت با هم ترکیب شده و در نهایت درخت تصمیم گیری نهایی برای اعتبارسنجی مشتریان بانک ساخته می شود.

با توجه به اینکه انتخاب تعداد بهینه خوشه ها از مسائل پیچیده می باشد، در مدل ترکیبی پیشنهادی می توان ابتدا تعداد خوشه را عدد ۲ در نظر گرفت و با توجه به نیاز و نظر کاربر، مرتبا تعداد آن ها را افزایش داد تا به بهترین تعداد خوشه ای دست یافت که درختان تصمیم گیری در آن خوشه ها بهترین درختان تصمیم گیری مبتنی بر معیارهای بهینگی باشند. البته در این مقاله تعداد خوشه عدد ۲ درنظر گرفته شده و از این روش برای انتخاب تعداد بهینه خوشه ها استفاده نمی شود.

به نظر می رسد بین درصد مشاهدات درست طبقه بندی شده و سایر معیارهای بهینگی درختان تصمیم گیری در برخی مواقع تضاد بوجود آید. به عبارت دیگر افزایش درصد مشاهدات درست طبقه بندی شده ممکن است باعث افزایش تعداد ویژگی های پیشگویی کننده انتخابی، تعداد برگ ها و اندازه درخت تصمیم گیری شود. می توان با روش هایی مثل هرس درخت تصمیم گیری و اعمال محدودیت هایی مثل مینیموم تعداد تراکنش در هر برگ به برقراری تعادل بین معیارهای بهینگی درختان تصمیم گیری پرداخت. ولی برای مقایسه درختان تصمیم گیری C4.5 نیز باید یک تعاملی بین ۴ معیار بهینگی درخت تصمیم گیری بوجود آید. ممکن است درخت تصمیم با دقت کمتر، دارای اندازه و تعداد برگ های کمتری در درخت تصمیم گیری نیز باشد. در صورتی که کاهش دقت نامحسوس باشد، با توجه به نظر کاربر درخت تصمیم گیری با دقت کمتر برای طبقه بندی مشتریان بانک ها انتخاب می شود. زیرا این درخت تصمیم گیری دارای اندازه و تعداد برگ های کمتری نسبت به درخت تصمیم دیگر است. در نهایت اینکه تصمیم گیری نهایی در انتخاب درخت تصمیم گیری C4.5 بهتر در هر خوشه به نظر کاربر یا کارشناس اعتبارسنجی بستگی دارد.

در چهار الگوریتم اول انتخاب ویژگی مبتنی بر الگوریتم ژنتیک مطرح در بخش ۵,۴ این مقاله، از روش گلدبرگ برای نمایش ژنتیکی کروموزوم ها استفاده می شود. هر کروموزوم نشان دهنده زیر مجموعه ویژگی ها است. هر ژن نماد یک ویژگی است. مقدار آن ژن برابر یک و صفر است که به ترتیب نشان دهنده وجود و عدم وجود ویژگی مورد نظر در زیر مجموعه ویژگی ها است. عملگر انتخاب از چرخ گردان برای انتخاب کروموزوم های برتر استفاده می کند. از عملگر تقاطع تک نقطه ای برای انجام عمل تقاطع استفاده می شود. بدین ترتیب که به طور تصادفی برشی بر روی دو کروموزوم ایجاد می شود و قسمت های برش خورده به صورت ضربدری یا اریب با هم جابجا می شوند. با انجام این عمل دو کروموزوم جدید بوجود می آید. عملگر جهش بدین صورت است که اگر مقدار یک ژن که به صورت تصادفی انتخاب می شود، صفر باشد، آن را تبدیل به یک می کند و اگر مقدار آن ژن یک باشد، آن را به صفر تبدیل می نماید. عملگر جایگزینی کروموزوم ها با کروموزوم های قبلی بر پایه شایستگی است. شرط توقف الگوریتم ژنتیک در اینجا تعداد نسل ها در نظر گرفته شد.

۵- آموزش، تست مدل

به منظور آموزش و تست مدل ترکیبی پیشنهادی پس از آماده سازی داده های اعتباری آلمان، تعداد ۶۹۰ تراکنش از ۱۰۰۰ تراکنش مورد استفاده قرار گرفت. تعداد خوشه ها در الگوریتم خوشه بندی

SimpleKmeans و عدد seed به ترتیب برابر دو و یک در نظر گرفته شد. پارامترهای الگوریتم ژنتیک در چهار الگوریتم انتخاب ویژگی ابتدایی در بخش ۵,۴ به قرار زیر است: نرخ تقاطع ۰,۹، نرخ جهش ۰,۰۱، تعداد نسل و جمعیت اولیه ۲۰ و عدد تصادفی seed برابر ۱ در نظر گرفته شد. از عدد اعتبارسنجی متقاطع ۱۰ برای آموزش و تست مدل استفاده شد. بدین ترتیب که ابتدا یک دهم اول داده ها برای تست استفاده می شود و بقیه برای آموزش الگوریتم انتخاب ویژگی یا درخت تصمیم گیری C4.5 بکار می رود. سپس یک دهم بعدی و به همین ترتیب ۱۰ بار این عمل صورت می گیرد و از نتایج این مراحل میانگین گرفته می شود. تعداد دسته ها و عدد seed و حد آستانه[۳۰] در الگوریتم انتخاب ویژگی با تابع ارزیاب Wrapper با طبقه کننده C4.5 به ترتیب برابر ۱۰ و ۱ و ۰,۰۱ است.

مقادیر پارامترهای الگوریتم انتخاب ویژگی مبتنی بر درخت تصمیم گیری ژنتیکی به صورت زیر است. استفاده از رویکرد اعتبارسنجی متقاطع با عدد ۱۰ در آموزش و تست درخت تصمیم گیری ژنتیکی، عملگر تقاطع: تصادفی استاندارد، عمل جهش: تصادفی استاندارد، درصد جایگزینی ژنوم یعنی درصد تعداد درختان بد که در نسل ها جایگزین می شوند، برابر ۰,۲۵، نرخ خطا برابر ۰,۹۵(زمانی که خطای طبقه بندی یک درخت از حد معین تعریف شده توسط این پارامتر بیشتر شود، از طبقه بندی تراکنش ها در مجموعه تست جلوگیری می شود تا منابع حفظ شود. با افزایش مقدار این پارامتر، سرعت تکامل[۳۱] در الگوریتم درخت تصمیم ژنتیکی افزایش می یابد.)، ترجیح قابلیت درخت تصمیم با دقت بالاتر بر درخت تصمیم کوچکتر، عدم تغییر پویا در ترجیح درختان تصمیم گیری با دقت بالاتر بر درخت تصمیم گیری کوچکتر در ابتدا و انتهای فرایند تکامل، نرخ تقاطع عدد ۰,۹۹، نرخ جهش عدد ۰,۰۱، تعداد نسل ها عدد 100، جمعیت اولیه عدد 100، عدد تصادفی seed برابر ۱۲۳۴۵۶۷۸۹.

مشخصات درخت تصمیم گیری C4.5 به قرار زیر است: ماتریس هزینه های طبقه بندی غلط در ساخت درخت تصمیم گیری C4.5 در نظر گرفته نشد، استفاده از هرس درخت، استفاده از فاکتور اطمینان[۳۲] 0.25 در فرایند ساخت درخت تصمیم گیری و تعیین مینیموم تعداد تراکنش با عدد ۲ در هر برگ درخت تصمیم گیری برای هرس درخت و تعیین اندازه و پیچیدگی درخت، عدم استفاده از جداکننده های دودویی، تعداد دسته ها عدد ۳(یعنی تعداد دو و یک دسته به ترتیب برای رشد و هرس درخت)، عدم استفاده از روش هرس خطای کاهش یافته، عدد تصادفی seed برابر ۱ برای تصادفی کردن داده در زمان استفاده از روش هرس خطای کاهش یافته، زیاد کردن زیر درخت، رویکرد اعتبارسنجی متقاطع در آموزش و تست درخت تصمیم گیری C4.5 با عدد ۱۰. بهتر است در ساخت درختان تصمیم گیری و مخصوصا اعتبارسنجی مشتریان بانک ها همواره بین میزان پیچیدگی و دقت طبقه بندی مدل های طبقه بندی یک تعادل برقرار باشد. کاهش فاکتور اطمینان منجر به افزایش هرس درخت می شود. افزایش مینیموم تعداد تراکنش در هر برگ باعث می شود تعداد بیشتری تراکنش در یک برگ قرار گیرد و در نتیجه تعداد برگ ها، تعداد ویژگی های پیشگو منتخب و اندازه درخت کاهش یافته، ولی احتمال کاهش دقت طبقه بندی وجود دارد. پس برقراری تعادل بین معیارهای بهینگی درختان تصمیم گیری برای اعتبارسنجی مشتریان بانک ها ضروری است که در این مقاله این موضوع لحاظ شده است.

۶- مقایسه درخت تصمیم گیری حاصل از مدل ترکیبی پیشنهادی با سایر درختان تصمیم گیری

تا این قسمت از مقاله به ارائه مدل ترکیبی پیشنهادی برای اعتبارسنجی مشتریان بانک پرداخته شد. این مدل به طور مختصر با توجه به شکل ۱ مراحل زیر را برای اعتبارسنجی مشتریان بانک ها انجام می دهد: ۱. خوشه بندی داده ها. ۲. انتخاب ویژگی ها توسط پنج الگوریتم انتخاب ویژگی. ۳. ساخت درختان تصمیم گیری C4.5 از هر یک از مجموعه ویژگی های انتخابی در هر خوشه. ۴. انتخاب بهترین درختان تصمیم گیری در هر خوشه مبتنی بر معیارهای بهینگی. ۵. ترکیب دو درخت

تصمیم گیری بهتر در هر خوشه و ساخت درخت تصمیم گیری نهایی برای اعتبارسنجی مشتریان بانک.

در این بخش به ارائه یافته ها و تحلیل نتایج حاصل از اجرای مدل ترکیبی پیشنهادی این مقاله پرداخته شده و این نتایج با نتایج سایر روش های ساخت درختان تصمیم گیری مطرح در این مقاله مقایسه می شود. بدین منظور از مجموعه داده های اعتباری آلمان که عملیات آماده سازی روی آن ها اعمال شد، استفاده می شود. نتایج جداول ۴ الی ۹ مبتنی بر خروجی نرم افزار وکا در ساخت درختان تصمیم‌گیری C4.5 است. تنها برای انتخاب ویژگی ها مبتنی بر درخت تصمیم گیری ژنتیکی از نرم افزار GATree استفاده شده است. جدول ۴ به ارائه نتایج حاصل از اجرای مدل ترکیبی پیشنهادی و ساخت درختان تصمیم گیری C4.5 در خوشه اول می پردازد.

تعداد ویژگی های حاصل از اجرای پنج الگوریتم انتخاب ویژگی مطرح در بخش ۵،۴ به ترتیب برابر است با: ۷، ۱۵، ۱۶، ۱۴ و ۷. همان طور که در جدول ۴ مشاهده می شود، درخت تصمیم گیری C4.5 با تابع ارزیاب "طبقه کننده C4.5" در انتخاب ویژگی، دارای بیشترین درصد مشاهدات درست طبقه بندی شده است. از طرفی این درخت بیشترین ویژگی پیشگویی کننده انتخابی را دارا می باشد. به نظر می رسد این درخت در خوشه اول بهترین درخت باشد؛ زیرا نسبت به بقیه درختان تصمیم گیری در خوشه اول دارای بالاترین دقت در طبقه بندی بوده و به میزان کمی دارای پیچیدگی بیشتر نسبت به برخی از درختان تصمیم گیری است. کمترین تعداد ویژگی پیشگویی کننده انتخابی، تعداد برگ ها و اندازه درخت، مربوط به درخت تصمیم گیری است که از الگوریتم درخت تصمیم گیری ژنتیکی برای انتخاب ویژگی ها بهره برده است. شاید دلیل این امر تعداد ویژگی های انتخابی کم برای ساخت درخت تصمیم گیری C4.5 است که از الگوریتم درخت تصمیم گیری ژنتیکی برای انتخاب ویژگی ها استفاده کرده است. اما این درخت دارای دقت کمتری در طبقه بندی مشتریان نسبت به اکثر درختان تصمیم گیری در جدول ۴ است.

کمترین میزان دقت طبقه بندی، مربوط به درخت تصمیم گیری C4.5 با تابع ارزیاب "سازگاری زیر مجموعه ویژگی ها با مقادیر ویژگی هدف" در انتخاب ویژگی است. به نظر می رسد بدترین درخت به لحاظ معیارهای بهینگی، این درخت باشد. زیرا دارای کمترین دقت در طبقه بندی و بیشترین تعداد برگ ها و اندازه درخت را در بین درختان تصمیم گیری جدول ۴ دارا است. جدول ۵ به ارائه نتایج حاصل از اجرای مدل ترکیبی پیشنهادی و ساخت درختان تصمیم گیری C4.5 در خوشه دوم می پردازد. تعداد ویژگی های حاصل از اجرای پنج الگوریتم انتخاب ویژگی مطرح در بخش ۵،۴ به ترتیب برابر است با: ۶، ۱۲، ۱۵، ۶ و ۹.

جدول ۴ - نتایج حاصل از اجرای الگوریتم های ساخت درخت تصمیم گیری C4.5 مبتنی بر مدل ترکیبی پیشنهادی در خوشه اول

ردیف	تابع ارزیاب انتخاب ویژگی مبتنی بر الگوریتم ژنتیک	کل مشاهدات	تعداد ویژگی های پیشگویی کننده انتخابی	تعداد مشاهدات درست طبقه بندی شده	درصد مشاهدات درست طبقه بندی شده	تعداد برگ ها	اندازه درخت	دقت کلاس مشتریان خوب	دقت کلاس مشتریان بد
۱	wrapperبا طبقه کننده C4.5	۴۴۵	۱۱	۳۷۳	٪۸۳,۸۲۰۲	۲۸	۴۲	۰,۸۲۹	۰,۸۶۲
۲	همبستگی بین ویژگی ها با هم و با ویژگی هدف	۴۴۵	۵	۳۵۳	٪۷۹,۳۲۵۸	۱۸	۲۶	۰,۸۰۹	۰,۷۶۱
۳	سازگاری زیر مجموعه ویژگی ها با مقادیر ویژگی هدف	۴۴۵	۱۳	۳۱۶	٪۷۱,۰۱۱۲	۴۳	۶۳	۰,۷۲۲	۰,۶۷۳
۴	طبقه کننده C4.5	۴۴۵	۱۴	۳۷۶	۸۴,۴۹۴۴	۳۶	۵۵	۰,۸۳	۰,۸۸۲
5	مبتنی بر درخت تصمیم گیری ژنتیکی	۴۴۵	۶	۳۴۳	٪۷۷,۰۷۸۷	۵	۸	۰,۷۳۵	۰,۹۷۱

جدول ۵ - نتایج حاصل از اجرای الگوریتم های ساخت درخت تصمیم گیری C4.5 مبتنی بر مدل ترکیبی پیشنهادی در خوشه دوم

ردیف	تابع ارزیاب انتخاب ویژگی مبتنی بر الگوریتم ژنتیک	کل مشاهدات	تعداد ویژگی های پیشگویی کننده انتخابی	تعداد مشاهدات درست طبقه بندی شده	درصد مشاهدات درست طبقه بندی شده	تعداد برگ ها	اندازه درخت	دقت کلاس مشتریان خوب	دقت کلاس مشتریان بد
۱	wrapperبا طبقه کنندهC4.5	۲۴۵	۳	۲,۶	٪۸۴,۰۸۱۶	۶	۹	۰,۸۵۶	۰,۸۱۲
۲	همبستگی بین ویژگی ها با هم و با ویژگی هدف	۲۴۵	۳	۲۰۶	٪۸۴,۰۸۱۶	۶	۹	۰,۸۵۶	۰,۸۱۲
۳	سازگاری زیر مجموعه ویژگی ها با مقادیر ویژگی هدف	۲۴۵	۹	۱۸۰	٪۷۳,۴۶۹۴	۱۹	۲۷	۰,۷۲	۰,۸
۴	طبقه کننده C4.5	۲۴۵	۹	۲۰۳	٪۸۲,۸۵۷۱	۱۹	۲۹	۰,۸۴۵	۰۷۹۸
5	مبتنی بر درخت تصمیم گیری ژنتیکی	۲۴۵	۳	۱۹۰	٪۷۷,۵۵۱	۷	۱۰	۰,۷۴۵	۰,۹۱۱

جدول ۶ - نتایج حاصل از اجرای الگوریتم های ساخت درخت تصمیم گیری C4.5 مبتنی بر مدل ترکیبی پیشنهادی

ردیف	الگوریتم درخت تصمیم گیری	کل مشاهدات	تعداد ویژگی های پیشگویی کننده انتخابی	تعداد مشاهدات درست طبقه بندی شده	درصد مشاهدات درست طبقه بندی شده	تعداد برگ ها	اندازه درخت	دقت کلاس مشتریان خوب	دقت کلاس مشتریان بد
۱	درخت تصمیم گیری حاصل از مدل ترکیبی پیشنهادی	۶۹۰	۱۵	۵۸۲	٪۸۴,۳۴۷۸	۴۲	۶۵	۰,۸۳۸۹	۰,۸۵۳۷

تعداد ویژگی های انتخابی در الگوریتم های انتخاب ویژگی با توابع ارزیاب "همبستگی بین ویژگی ها با هم و با ویژگی هدف" و "Wrapper با طبقه کننده C4.5" کم است. از طرف دیگر پیچیدگی درختان تصمیم گیری C4.5 که از این دو روش برای انتخاب ویژگی ها استفاده کرده اند، نسبت به سایر درختان تصمیم گیری در جدول ۵ کمتر شده است. زیرا دارای تعداد ویژگی پیشگویی کننده انتخابی، تعداد برگ ها و اندازه درخت کمتری هستند. بیشترین میزان "درصد مشاهدات درست طبقه بندی شده" نیز مربوط به این دو درخت است. درخت تصمیم گیری با تابع ارزیاب "سازگاری زیر مجموعه ویژگی ها با مقادیر ویژگی هدف" دارای کمترین میزان دقت در طبقه بندی مشتریان بوده و پیچیدگی آن نیز نسبتا بالا است. درخت تصمیم گیری حاصل از الگوریتم انتخاب ویژگی مبتنی بر درخت تصمیم گیری ژنتیکی دارای دقت طبقه بندی کمتری نسبت به ۳ درخت تصمیم گیری جدول ۵ است. پیچیدگی آن نیز در حد قابل قبولی پایین است. ولی به دلیل کم بودن میزان دقت طبقه بندی، برتری نسبت به دو درخت تصمیم گیری با دقت بالا در جدول ۵ ندارد. بهترین درخت در خوشه دوم، درخت تصمیم گیری است که تابع ارزیاب آن در انتخاب ویژگی ها برابر همبستگی بین ویژگی ها با هم و با ویژگی هدف یا Wrapper با طبقه کنندهC4.5 است؛ زیرا هر دو درخت نتایج یکسانی را ارائه می دهند. این درختان دارای بالاترین دقت طبقه بندی در بین درختان تصمیم گیری خوشه دوم بوده و به لحاظ پیچیدگی نسبت به بسیاری از درختان تصمیم گیری بهتر هستند.

بر طبق مدل ترکیبی پیشنهادی مقاله در اعتبارسنجی مشتریان بانک ، بعد از اینکه در هر خوشه درختان تصمیم گیری ساخته شدند، نوبت به انتخاب بهترین درخت تصمیم گیری مبتنی بر معیارهای بهینگی می رسد. این کار در بالا انجام شد. بعد از انتخاب بهترین درختان تصمیم گیری در هر خوشه نوبت به ترکیب دو درخت تصمیم گیری و ساخت درخت تصمیم گیری نهایی برای اعتبارسنجی مشتریان بانک می رسد. نتایج حاصل از ساخت درخت تصمیم گیری توسط مدل

ترکیبی پیشنهادی این مقاله در مجموعه داده های اعتباری آلمان به صورت جدول ۶ است.

"ویژگی های پیشگویی کننده انتخابی" بهترین درخت تصمیم گیری خوشه اول در "ویژگی های پیشگویی کننده انتخابی" بهترین درخت تصمیم گیری خوشه دوم موجود است. اگر ویژگی "نوع خوشه" به این مجموعه ویژگی اضافه شود، تعداد کل ویژگی های پیشگویی کننده انتخابی درخت تصمیم گیری مدل ترکیبی پیشنهادی برابر با ۱۵ است. تعداد مشاهدات درست طبقه بندی شده برابر با مجموع تعداد مشاهدات درست طبقه بندی شده در درختان تصمیم گیری منتخب در دو خوشه است. تعداد برگ ها در درخت تصمیم گیری مدل ترکیبی پیشنهادی برابر با مجموع تعداد برگ های درختان تصمیم گیری منتخب در هر دو خوشه است. اندازه درخت در این مدل برابر مجموع اندازه درخت در درختان تصمیم گیری دو خوشه به علاوه گره با ویژگی "نوع خوشه" است. این ویژگی در ابتدای درخت تصمیم گیری مدل ترکیبی پیشنهادی به تعیین نوع خوشه به منظور طبقه بندی توسط درخت تصمیم گیری منتخب در هر خوشه می پردازد. ویژگی "نوع خوشه" یک ویژگی اسمی می باشد و در مدل پیشنهادی دارای مقادیر "خوشه اول" و "خوشه دوم" است.

می توان به ساخت درختان تصمیم گیری به وسیله مدل ترکیبی پیشنهادی پرداخت. تنها با این تفاوت که در ابتدا از خوشه بندی برای تفکیک مشتریان به دو خوشه اول و دوم استفاده نکرد. جدول ۷ نتایج حاصل از اجرای مدل ترکیبی پیشنهادی بدون خوشه بندی را نشان می دهد.

همان طور که در جدول ۷ مشاهده می شود، درختان تصمیم گیری با توابع ارزیاب Wrapper با طبقه کننده C4.5 و "مبتنی بر درخت تصمیم گیری ژنتیکی در انتخاب ویژگی" بهترین درختان به ترتیب در معیارهای دقت و پیچیدگی هستند. بهترین درخت تصمیم گیری در جدول ۷، درخت تصمیم گیری با تابع ارزیاب Wrapper با طبقه کننده C4.5 است. زیرا بالاترین دقت را در طبقه بندی و اعتبارسنجی مشتریان دارد. همچنین پیچیدگی آن نسبت به بقیه درختان تصمیم گیری به جز درخت تصمیم گیری با انتخاب ویژگی ها مبتنی بر درخت تصمیم گیری ژنتیکی کمتر است. با مقایسه نتایج جداول ۶ و ۷ مشاهده می شود که خوشه بندی مشتریان قبل از ساخت درختان تصمیم گیری در مدل ترکیبی پیشنهادی نتایج بهتری را در معیار درصد مشاهدات درست طبقه بندی شده در مجموعه داده اعتباری آلمان ارائه داد. در صورتی که نیاز به تصمیم گیری منعطف، ساده و بدون پیچیدگی در اعتبارسنجی است، از درخت تصمیم گیری استفاده شود که در انتخاب ویژگی ها از تابع ارزیاب Wrapper با طبقه کننده C4.5 استفاده کرده است و از قبل عمل خوشه بندی برای ساخت آن صورت نگرفته است.

جدول ۷ - نتایج حاصل از اجرای مدل ترکیبی پیشنهادی بدون استفاده از خوشه بندی در ساخت درخت تصمیم گیری C4.5

ردیف	تابع ارزیاب انتخاب ویژگی مبتنی بر الگوریتم ژنتیک	کل مشاهدات	تعداد ویژگی های پیشگویی کننده انتخابی	تعداد مشاهدات درست طبقه بندی شده	درصد مشاهدات درست طبقه بندی شده	تعداد برگ ها	اندازه درخت	دقت کلاس مشتریان خوب	دقت کلاس مشتریان بد
۱	wrapperبا طبقه کننده C4.5	۶۹۰	۶	۵۷۰	٪ ۸۲,۶۰۸۷	۱۴	۲۱	۰,۸۲۶	۰,۸۲۵
۲	همبستگی بین ویژگی ها با هم و با ویژگی هدف	۶۹۰	۶	۵۵۳	٪ ۸۰,۱۴۴۹	۲۲	۳۴	۰,۸۰۹	۰,۷۸۴
۳	سازگاری زیر مجموعه ویژگی ها با مقادیر ویژگی هدف	۶۹۰	۱۴	۴۸۹	٪ ۷۰,۸۶۹۶	۶۸	۱۰۱	۰,۷۲۳	۰,۶۶۷
۴	طبقه کننده C4.5	۶۹۰	۱۵	۵۶۷	٪ ۸۲,۱۷۳۹	۴۹	۷۳	۰,۸۱۵	۰,۸۳۹
5	مبتنی بر درخت تصمیم گیری ژنتیکی	۶۹۰	۴	۵۶۲	٪ ۸۱,۴۴۹۳	۱۱	۱۸	۰,۸۲۲	۰,۷۹۸

می توان ویژگی "نوع خوشه" را در ساخت درخت تصمیم گیری در نظر گرفت. برای این کار به مجموعه داده اعتباری آلمان یک ویژگی پیشگویی کننده با نام ویژگی "نوع خوشه" اضافه می شود که توسط آن نوع خوشه هر تراکنش مشتری تعیین شده است. پارامترهای خوشه بندی در اینجا همانند مدل ترکیبی پیشنهادی است. جدول ۸ نتایج حاصل از اجرای این الگوریتم را نشان می دهد. الگوریتم ساخت درختان تصمیم گیری در این روش همانند الگوریتم ساخت درختان تصمیم گیری در مدل ترکیبی پیشنهادی است. مقادیر پارامترهای الگوریتم ها در این روش همانند الگوریتم های ساخت درختان تصمیم گیری در مدل ترکیبی پیشنهادی است.

همان طور که در جدول ۸ مشاهده می شود، درخت تصمیم گیری با تابع ارزیاب Wrapper با طبقه کننده C4.5 دارای بالاترین میزان دقت طبقه بندی است. کمترین پیچیدگی مربوط به درخت تصمیم گیری است که در انتخاب ویژگی ها از درخت تصمیم گیری ژنتیکی استفاده کرده است. به نظر می رسد بهترین درخت تصمیم گیری در جدول ۸، درخت تصمیم گیری با تابع ارزیاب Wrapper با طبقه کننده C4.5 باشد. زیرا دارای دقت قابل توجه بالا در طبقه بندی مشتریان اعتبارسنجی نسبت به سایر درختان تصمیم گیری جدول ۸ است. به گونه ای که شاید بتوان از پیچیدگی نسبی این درخت نسبت به برخی از درختان تصمیم گیری جدول ۸ چشم پوشی کرد. نکته قابل ذکر در اینجا این است که ویژگی "نوع خوشه" در درخت تصمیم گیری با تابع ارزیاب Wrapper با طبقه کننده C4.5 به عنوان یک ویژگی پیشگویی کننده وجود دارد. با مقایسه نتایج جداول ۶ و ۸ مشاهده می شود که درخت تصمیم گیری C4.5 با تابع ارزیاب Wrapper با طبقه کننده C4.5 دارای دقت بالا و پیچیدگی کمتر نسبت به بهترین درخت تصمیم گیری C4.5حاصل از مدل ترکیبی پیشنهادی است.

برای اینکه بتوان درخت تصمیم گیری مدل ترکیبی پیشنهادی را با درختان تصمیم گیری دیگری مقایسه کرد که در آن ها از خوشه بندی و انتخاب ویژگی ها استفاده نشده است، نتایج دو درخت تصمیم گیری C4.5 در جدول ۹ آمده است.

جدول ۹ به ارائه نتایج اجرای دو نوع دیگر الگوریتم درخت تصمیم گیری C4.5 و مقایسه آن ها با درخت تصمیم گیری حاصل از مدل ترکیبی پیشنهادی می پردازد. در ردیف ۱ از جدول ۹ نتایج الگوریتم درخت تصمیم گیری C4.5 ارائه شده است که در آن از تکنیک های خوشه بندی و انتخاب ویژگی ها استفاده نشده است. این درخت دارای دقت طبقه بندی کمتری نسبت به درخت تصمیم گیری حاصل از مدل ترکیبی پیشنهادی است؛ ولی تعداد برگ ها، اندازه درخت و تعداد ویژگی های پیشگویی کننده انتخابی آن کمتر است. در نتیجه پیچیدگی کمتری دارد. در ردیف ۲ از جدول ۹ نتایج درخت تصمیم گیری C4.5 نشان داده شده است که همانند درخت ردیف ۱ از خوشه بندی و انتخاب ویژگی ها استفاده نکرده ولی ویژگی "نوع خوشه" در مجموعه ویژگی ها برای ساخت این درخت وجود دارد. این درخت تصمیم گیری دارای دقت طبقه بندی، تعداد ویژگی های پیشگویی کننده انتخابی، تعداد برگ ها و اندازه درخت کمتری نسبت به درخت تصمیم گیری مدل ترکیبی پیشنهادی است.

جدول ۸ – نتایج حاصل از اجرای الگوریتم ساخت درخت تصمیم گیری C4.5 با در نظر گرفتن ویژگی "نوع خوشه" در مجموعه ویژگی ها

ردیف	تابع ارزیاب انتخاب ویژگی مبتنی بر الگوریتم ژنتیک	کل مشاهدات	تعداد ویژگی های پیشگویی کننده انتخابی	تعداد مشاهدات درست طبقه بندی شده	درصد مشاهدات درست طبقه بندی شده	تعداد برگ ها	اندازه درخت	دقت کلاس مشتریان خوب	دقت کلاس مشتریان بد
۱	wrapperبا طبقه کننده C4.5	۶۹۰	۱۳	۵۸۴	٪ ۸۴,۶۳۷۷	۳۷	۵۹	۰,۸۳۷	۰,۸۶۹
۲	همبستگی بین ویژگی ها با هم و با ویژگی هدف	۶۹۰	۴	۵۵۶	٪ ۸۰,۵۷۹۷	۱۱	۱۸	۰,۸۱۷	۰,۷۸۲
۳	سازگاری زیر مجموعه ویژگی ها با مقادیر ویژگی هدف	۶۹۰	۱۳	۴۸۲	٪ ۶۹,۸۵۵۱	۶۲	۹۳	۰,۷۳۹	۰,۶۱۳

۴	طبقه کننده C4.5	۶۹۰	۱۵	۵۶۴	٪ ۸۱,۷۳۹۱	۴۸	۷۶	۰,۸۱۹	۰,۸۱۴
۵	مبتنی بر درخت تصمیم گیری ژنتیکی	۶۹۰	۴	۵۴۴	٪ ۷۸,۸۴۰۶	۹	۱۳	۰,۷۴۷	۱

جدول ۹ – نتایج حاصل از بکارگیری الگوریتم های درخت تصمیم گیری C4.5

ردیف	الگوریتم درخت تصمیم گیری	کل مشاهدات	تعداد ویژگی های پیشگویی کننده انتخابی	تعداد مشاهدات درست طبقه بندی شده	درصد مشاهدات درست طبقه بندی شده	تعداد برگ ها	اندازه درخت	دقت کلاس مشتریان خوب	دقت کلاس مشتریان بد
۱	درخت تصمیم گیری C4.5	۶۹۰	۱۴	۵۷۳	٪ ۸۳,۰۴۳۵	۴۰	۶۰	۰,۸۲۶	۰,۸۴۱
۲	درخت تصمیم گیری C4.5 با ویژگی "نوع خوشه"	۶۹۰	۱۲	۵۷۲	٪ ۸۲,۸۹۸۶	۳۶	۵۶	۰,۸۳۱	۰,۸۲۴

در این بخش از مقاله به ارائه نتایج حاصل از اجرای مدل ترکیبی پیشنهادی مقاله و مقایسه آن با برخی درختان تصمیم گیری C4.5 پرداخته شد. در انتخاب بهترین درختان تصمیم گیری در هر خوشه نظرات کارشناسان اعتبارسنجی را می توان لحاظ کرد. همچنین پیچیدگی طبقه بندی و هزینه طبقه بندی غلط مشتریان اعتبارسنجی در انتخاب بهترین درختان تصمیم گیری مهم هستند.

۷- نتیجه گیری

بانک ها به منظور اعطای تسهیلات اعتباری به مشتریان خود نیازمند اعتبارسنجی آن ها هستند. تکنیک های متنوعی مثل درخت تصمیم گیری C4.5 می تواند به بانک ها در شناسایی مشتریان خوب و بد به لحاظ اعتباری کمک کند. انتخاب درخت تصمیم گیری بهتر که بتواند نسبت به سایر درختان تصمیم گیری به طبقه بندی مشتریان با دقت بالا و پیچیدگی کمتر بپردازد، می تواند باعث شود تا کارشناسان اعتبارسنجی بتوانند با صرف زمان و هزینه کم و افزایش رضایت مشتری اعتبارسنجی به طبقه بندی مشتریان خود بپردازند. در این مقاله به ارائه مدل جدیدی پرداخته شد که با استفاده از تکنیک های الگوریتم ژنتیک، انتخاب ویژگی ها و خوشه بندی به ساخت درختان تصمیم گیری C4.5 بپردازد. برای آزمون مدل از مجموعه داده اعتباری آلمان استفاده شد. در نهایت مبتنی بر معیارهای بهینگی درختان تصمیم گیری مطرح در این مقاله، بهترین درختان تصمیم گیری در هر خوشه انتخاب شده و با ترکیب آن ها درخت تصمیم گیری حاصل از فرایند ساخت و آزمون مدل ترکیبی پیشنهادی برای اعتبارسنجی مشتریان بانک ها ایجاد شد. نتایج حاصل از اجرای الگوریتم مدل ترکیبی پیشنهادی نشان داد که دقت طبقه بندی درخت تصمیم گیری حاصل از آن نسبت به بسیاری از درختان تصمیم گیری مقایسه شده در این مقاله بیشتر بود. تنها درخت تصمیم گیری با تابع ارزیاب Wrapper با طبقه کننده C4.5 و مبتنی بر الگوریتم ژنتیک در انتخاب ویژگی ها که ویژگی "نوع خوشه" را در ساخت خود لحاظ کرد، دارای بیشترین دقت طبقه بندی بوده و پیچیدگی آن نیز از درخت تصمیم گیری مدل ترکیبی پیشنهادی کمتر بود. پیچدگی درخت تصمیم گیری مدل ترکیبی پیشنهادی نسبت به برخی از درختان تصمیم گیری بیشتر بود؛ ولی دقت بالای طبقه بندی آن باعث شد این درخت یک برتری نسبی نسبت به آن ها در اعتبارسنجی مشتریان داشته باشد. با توجه به نکات بالا می توان از مدل ترکیبی پیشنهادی برای ساخت درختان تصمیم گیری بهنیه در اعتبارسنجی مشتریان بانک ها استفاده نمود. همچنین می توان هزینه طبقه بندی غلط مشتریان خوب و بد را در ساخت درختان تصمیم گیری لحاظ کرد که این شاخص در این مقاله، در نرم افزار وکا برای دو نوع مشتری یکسان لحاظ شد.

فهرست منابع

۱) اداره مطالعات و کنترل ریسک بانک تجارت(۱۳۸۶). "مدیریت ریسک در بانکداری". چاپ اول. تهران: موسسه انتشارات و چاپ دانشگاه تهران.

۲) نادعلی, ا., & خان بابایی, م.(۱۳۸۷)، "بکارگیری تکنیک های درخت تصمیم و الگوریتم ژنتیک جهت اعتبارسنجی مشتریان بانک ها در یک سیستم پشتیبانی تصمیم گیری", دومین کنفرانس ملی داده کاوی, تهران: دانشگاه صنعتی امیرکبیر, صفحه 8,6..

3) Abdou, H., & Pointon, J,(2008),"Neural nets versus conventional techniques in credit scoringin Egyptian banking",Expert Systems with Applications,Vol.xxx,pp.1.

4) Aitkenhead, M. J,(2008),"A co-evolving decision tree classification method",Expert Systems with Applications,Vol.34,pp.2-3.

5) Bala, J., Huang, J., Vafaie, H., DeJong, K., & Wechsler, H, (1995)," Hybrid Learning Using Genetic Algorithms and Decision Trees for Pattern Classification",IJCAI,Montreal: IJCAI conference,pp. 1,2,4.

6) Carvalho, D. R., & Freitas, A. A. (2002)"A genetic-algorithm for discovering small-disjunct rules in data mining", Applied Soft Computing ,Vol.2,pp. 2-8,12.

7) Carvalho, D. R., & Freitas, A. A, (2004)," A hybrid decision tree/genetic algorithm method for data mining",Information Sciences,Vol.163,pp.1-18.

8) D'heygere, T., Goethals, P. L., & Pauw, N. D. ,(2006),"Genetic algorithms for optimisation of predictive ecosystems models based on decision trees and neural networks", Ecological Modelling,Vol.195,pp.1-5.

9) D'heygere, T., Goethals, P. L., & Pauw, N. D,(2003),"Use of genetic algorithms to select input variables in decision tree models for the prediction of benthic macroinvertebrates",Ecological Modelling,Vol.160,pp.1-8.

10) Dehuri, S., Patnaik, S., Ghosh, A., & Mall, R., (2008),"Application of elitist multi-objective genetic algorithm for classification rule generation". Applied Soft Computing ,Vol.8,pp.1,2,3,5.

11) Gray, J. B., & Fan, G. (2008),"Classification tree analysis usingTARGET". Computational Statistics & Data Analysis ,Vol.52,pp.1-3.

12) Grupe, F. H., & Jooste, S,(2004), "Genetic Algorithms A Business Perspective",Information Management & Computer Security,Vol.12,No.3,pp.1-4.

13) Guyon, I., & Elisseeff, A,(2003),"An Introduction to Variable and Feature Selection",Journal of Machine Learning Research,Vol.3,pp.2,4,9.

14) Hall, M. A,(1999),"Corrolation Based Feature Selection for Machine Learning." Ph.D.,University of Waikato.

15) Hsu, P. L., Lai, R., Chiu, C. C., & Hsu, C. I. ,"(2003),The hybrid of association rule algorithms and genetic algorithms for tree induction:an example of predicting the student course performance". ExpertSystemswithApplications ,Vol.25,pp.1,2,4,5,6,7,11.

16) Huang, C. L., Chen, M. C., & Wang, C. J,"(2007),Credit Scoring with a data mining Approach Based on Support Vector Machines", Expert Systems with Apllications,Vol.33,pp.1-3.

17) Huang, M., Gong, J., Shi, Z., Liu, C., & Zhang, L. ,(2007),"Genetic algorithm-based decision tree classifier for remote sensing mapping with SPOT-5 data in the HongShiMao watershed of the loess plateau,China". Neural Comput & Applic, Available at: www.springer.com,pp.1-3.

18) http://archive.ics.uci.edu/ml/datasets.html.

19) Kennedy, R. L., Lee, Y., Roy, B. V., Reed, C. D., & Lippmann, R. P,(1998),"Solving Data Mining Problems through Pattern Recognition", S.l.,Prentice Hall.

20) Kim, E., Kim, W., & Lee, Y,(2002),"Combination of multiple classifiers for the customer's purchase behavior prediction",Decision Support Systems,Vol.34,pp.2-8.

21) Kim, M. J., & Han, I.,(2003),"The discovery of experts' decision rules from qualitative bankruptcy data using genetic algorithms". Expert Systems with Applications ,Vol.25,pp.1-5,8.

22) Kim, Y. S., & Sohn, S. Y,(2004),"Managing loan customers using misclassification patterns of credit scoring model", Expert Systems with Applications,Vol.26,pp.1-3.

23) Larose, D. T,(2005),"Discovering Knowledge in Data, an Introduction to Data Mining", New Jersey: WILEY.

24) Lavrac, N., Gamberger, D., & Turney, P,(1995),"Cost-Sensitive Classification: Empirical Evaluation of a Hybrid Genetic Decision Tree Induction Algorithm",Journal of Artificial Intelligence Research,Vol.2,pp.1,2,4,5.
25) Lee, T. S., & Chen, I. F,(2005),"A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines",Expert Systems with Applications,Vol.28,pp.1,2,5,6,7,8.
26) Lee, T. S., Chiu, C. C., Lu, C. J., & Chen, I. F,(2002),"Credit scoring using the hybrid neural discriminant technique",Expert Systems with Applications,Vol.23,pp.1,5,6,8.
27) Liu, H. H., & Ong, C. S,(2008),"Variable selection in clustering for marketing segmentation using genetic algorithms",Expert Systems with Applications,Vol.34,pp.1,3,4,5,6.
28) Martinez-Otzeta, J. M., Sierra, B., Lazkano, E., & Astigarraga, A., (2006),"Classifier hierarchy learning by means of genetic algorithms". Pattern Recognition Letters , Vol.27,pp.1,2,3,5,6.
29) Nanni, L., & Lumini, A,(2009),"An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring", Expert Systems with Applications,Vol.36, pp.1-4.
30) Olson, D., & Shi, Y.,(2007),"Introduction to Business Data Mining". Singapore: McGraw Hill Education.
31) Ong, C. S., Huang, J. J., & Tzeng, G. H,(2005),"Building credit scoring models using genetic programming",Expert Systems with Applications,Vol.29,pp.1-3.
32) PAL, N. R., NAND, S., & Kundu, M. K,(1998),"Self-crossover-a new genetic operator and its application to feature selection",International Journal of Systems Science,Vol.29,No.2,pp.2,4,5,6.
33) Papagelis, A., & Kalles, D,(n.d.),"Breeding Decision Trees Using Evolutionary Techniques",from website www.siteceer.com,pp.1-7.
34) Sabzevari, H., Soleymani, M., & Noorbakhsh, E,(n.d.),"A comparison between statistical and Data Mining methods for credit scoring in case of limited available data",s.n., pp.1-7.
35) SALAPPA, A., DOUMPOS, M., & ZOPOUNIDIS, C,(2007),"Feature selection algorithms in classification problems:an experimental evaluation",Optimization Methods and Software,Vol.22,No.1,pp.2-5.
36) Sorensen, K., & Janssens, G. K,(2003),"Data mining with genetic algorithms on binary trees",European Journal of Operational ResearchVol.151,pp.2,10.
37) Susac, M. Z., Sarlija, N., & Bensic, M,(n.d.),"Small Business Credit Scoring: A Comparison of Logistic Regression, Neural Network, and Decision Tree Models",s.n,pp.1-4.
38) Tan, F., Fu, X., Zhang, Y., & Bourgeois, A. G,(2008),"A genetic algorithm-based method for feature sub set selection",Soft Comput,form www.springer.com,pp.1,3,4,5,6.
39) Thomas, L. C,(2000),"A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers",International Journal of Forecasting,Vol.16,pp.2-4.
40) Tsang, C. H., Kwong, S., & Wang, H,(2007),"Genetic-fuzzy rule mining approach and evaluation of feature selection techniques for anomaly intrusion detection",Pattern Recognition,Vol.40,pp.7,10,13.
41) WANG, Y. Y., & LI, J,(2008),"Feature-selection ability of the decision-tree algorithm and the impact of feature-selection/extraction on decision-tree results based on hyperspectral data",International Journal of Remote Sensing,Vol.22,No.10,pp.4,6,7.
42) Xu, X., Zhou, C., & Wang, Z,(2008),"Credit scoring algorithm based on link analysis ranking with support vector machine",Expert Systems with Applications,Vol.xxx,pp.6.
43) Zhang, Y., & Bhattacharyya, S. ,"Genetic programming in classifying large-scale data:an ensemble method". Information Sciences , Vol.163, pp.2,6.

پیوست ها

پیوست ۱- مقادیر ویژگی های داده های اعتبارسنجی آلمان

ویژگی ها	نوع ویژگی ها	مقادیر ویژگی ها
وضعیت چک	اسمی	<0 ; >=200 ; 0<=X<200 ; no checking
مدت زمان	عددی	عددی
سابقه اعتبار	اسمی	all paid ; critical/other existing credit ; delayed previously ; existing paid ; no credits/all paid
هدف	اسمی	کسب و کار ; اسباب خانه ; آموزش ; مبلمان و تجهیزات ; ماشین جدید ; موارد دیگر ; رادیو تلویزیون ; تعمیرات ; آموزش مجدد ; ماشین دست دوم
مقدار اعتبار	عددی	عددی
وضعیت پس انداز	اسمی	<100 ; >=1000 ; 100<=X<500 ; 500<=X<1000 ; عدم پس انداز مشخص
سابقه کار	اسمی	<1 ; >=7 ; 1<=X<4 ; 4<=X<7 ; بیکار
تعداد اقساط	عددی	عددی
وضعیت شخصی و جنسیت	اسمی	female div/dep/mar ; male div/sep ; male mar/wid ; male single
طرف های دیگر	اسمی	متقاضی مشترک ; نه
محل اقامت فعلی	عددی	عددی
اموال و دارایی ها	اسمی	ماشین ; بیمه زندگی ; اموال ناشناخته ; املاک و ساختمان
سن	عددی	عددی
برنامه های پرداختی دیگر	اسمی	بانک ; نه ; موجودی
مسکن	اسمی	رایگان ; مالک ; اجاره
وضعیت اعتباری موجود	عددی	عددی
شغل	اسمی	high qualif/self emp/mgmt ; unemp/unskilled non res ; ماهر ; مقیم غیر ماهر
تعداد عائله	عددی	عددی
مالکیت تلفن	اسمی	نه ; بلی
کارگر خارجی	اسمی	بلی
طبقه(کلاس)	اسمی	بد ; خوب

یادداشت‌ها

[1] Credit Scoring
[2] Fisher
[3] David Durand
[4] Recursive
[5] Greedy
[6] Quinlan
[7] Information Gain
[8] Selected Predictive Attributes
[9] Correctly Classified Instances
[10] German Credit Data
[11] WEKA
[12] Visualization
[13] Discretization
[14] Merge Values
[15] Turning
[16] Ball Commitee
[17] Thomas
[18] Brill
[19] Character
[20] Capital
[21] Collateral
[22] Capacity
[23] Condition
[24] Entropy
[25] Bushiness
[26] Training Observations
[27] Evaluator Function
[28] Correlation Based Feature Selection
[29] John Holland
[30] Threshold
[31] Evolution
[32] Confidence Factor